3

THE GENBANK SEQUENCE DATABASE

Ilene Karsch-Mizrachi

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland

B. F. Francis Ouellette

Centre for Molecular Medicine and Therapeutics
Children's and Women's Health Centre of British Columbia
University of British Columbia
Vancouver, British Columbia

INTRODUCTION

Primary protein and nucleic acid sequence databases are so pervasive to our way of thinking in molecular biology that few of us stop to wonder how these ubiquitous tools are built. Understanding how the these databases are put together will allow us to move forward in our understanding of biology and in fully harvesting the abstracted information present in these records.

GenBank, the National Institutes of Health (NIH) genetic sequence database, is an annotated collection of all publicly available nucleotide and protein sequences. The records within GenBank represent, in most cases, single, contiguous stretches of DNA or RNA with annotations. GenBank files are grouped into divisions; some of these divisions are phylogenetically based, whereas others are based on the technical approach that was used to generate the sequence information. Presently, all records in GenBank are generated from direct submissions to the DNA sequence

databases from the original authors, who volunteer their records to make the data publicly available or do so as part of the publication process. GenBank, which is built by the National Center for Biotechnology Information (NCBI), is part of the International Nucleotide Sequence Database Collaboration, along with its two partners, the DNA Data Bank of Japan (DDBJ, Mishima, Japan) and the European Molecular Biology Laboratory (EMBL) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK). All three centers provide separate points of data submission, yet all three centers exchange this information daily, making the same database (albeit in slightly different format and with different information systems) available to the community at-large.

This chapter describes how the GenBank database is structured, how it fits into the realm of the protein databases, and how its various components are interpreted by database users. Numerous works have dealt with the topic of sequence databases (Bairoch and Apweiller, 2000; Baker et al., 2000; Barker et al., 2000; Benson et al., 2000; Mewes et al., 2000; Tateno et al., 1997). These publications emphasize the great rate at which the databases have grown, and they suggest various ways of utilizing such vast biological resources. From a practical scientific point of view, as well as from a historical perspective, the sequence data have been separated into protein and nucleotide databases. The nucleotides are the primary entry points to the databases for both protein and nucleotide sequences, and there appears to be a migration toward having the nucleotide databases also involved in ''managing'' the protein data sets, as will be illustrated below. This is not a surprising development, since submitters are encouraged to provide annotation for the coding sequence (CDS) feature, the feature that tells how a translation product is produced. This trend toward the comanagement of protein and nucleotide sequences is apparent from the nucleotide sequences available through Entrez (cf. Chapter 7) as well as with GenBank and the formatting of records in the GenPept format. It is also apparent at EBI, where SWISS-PROT and TREMBL are being comanaged along with EMBL nucleotide databases. Nonetheless, the beginnings of each database set are distinct. Also implicit in the discussion of this chapter is the underlying data model described in Chapter 2.

Historically, the protein databases preceded the nucleotide databases. In the early 1960s, Dayhoff and colleagues collected all of the protein sequences known at that time; these sequences were catalogued as the Atlas of Protein Sequences and Structures (Dayhoff et al., 1965). This *printed* book laid the foundation for the resources that the entire bioinformatics community now depends on for day-to-day work in computational biology. A data set, which in 1965 could easily reside on a single floppy disk (although these did not exist then), represented years of work from a small group of people. Today, this amount of data can be generated in a fraction of a day. The advent of the DNA sequence databases in 1982, initiated by EMBL, led to the next phase, that of the explosion in database sequence information. Joined shortly thereafter by GenBank (then managed by the Los Alamos National Laboratory), both centers were contributing to the input activity, which consisted mainly of transcribing what was published in the printed journals to an electronic format more appropriate for use with computers. The DNA Data Bank of Japan (DDBJ) joined the data-collecting collaboration a few years later. In 1988, following a meeting of these three groups (now referred to as the International Nucleotide Sequence Database Collaboration), there was an agreement to use a common format for data elements within a unit record and to have each database update only the records that

were directly submitted to it. Now, all three centers are collecting direct submissions and distributing them so that each center has copies of all of the sequences, meaning that they can act as a primary distribution center for these sequences. However, each record is owned by the database that created it and can only be updated by that database, preventing ''update clashes'' that are bound to occur when any database can update any record.

PRIMARY AND SECONDARY DATABASES

Although this chapter is about the GenBank nucleotide database, GenBank is just one member of a community of databases that includes three important protein databases: SWISS-PROT, the Protein Information Resource (PIR), and the Protein DataBank (PDB). PDB, the database of nucleic acid and protein structures, is described in Chapter 5. SWISS-PROT and PIR can be considered secondary databases, curated databases that add value to what is already present in the primary databases. Both SWISS-PROT and PIR take the majority of their protein sequences from nucleotide databases. A small proportion of SWISS-PROT sequence data is submitted directly or enters through a journal-scanning effort, in which the sequence is (quite literally) taken directly from the published literature. This process, for both SWISS-PROT and PIR, has been described in detail elsewhere (Bairoch and Apweiller, 2000; Barker et al., 2000.)

As alluded to above, there is an important distinction between primary (archival) and secondary (curated) databases. The most important contribution that the sequence databases make to the scientific community is making the sequences themselves accessible. The primary databases represent experimental results (with some interpretation) but are not a curated review. Curated reviews are found in the secondary databases. GenBank nucleotide sequence records are derived from the sequencing of a biological molecule that exists in a test tube, somewhere in a lab. They do not represent sequences that are a consensus of a population, nor do they represent some other computer-generated string of letters. This framework has consequences in the interpretation of sequence analysis. In most cases, all a researcher will need is a given sequence. Each such DNA and RNA sequence will be annotated to describe the analysis from experimental results that indicate why that sequence was determined in the first place. One common type of annotation on a DNA sequence record is the coding sequence (CDS). A great majority of the protein sequences have not been experimentally determined, which may have downstream implications when analyses are performed. For example, the assignment of a product name or function qualifier based on a subjective interpretation of a similarity analysis can be very useful, but it can sometimes be misleading. Therefore, the DNA, RNA, or protein sequences are the ''computable'' items to be analyzed and represent the most valuable component of the primary databases.

FORMAT VS. CONTENT: COMPUTERS VS. HUMANS

Database records are used to hold raw sequence data as well as an array of ancillary annotations. In a survey of the various database formats, we can observe that, although different sets of rules are applied, it is still possible in many cases to inter-

change data among formats. The best format for a human to read may not be the most efficient for a computer program. The simplicity of the flat file, which does lend itself to simple tools that are available to all, is in great part responsible for the popularity of the EMBL and GenBank flatfile formats. In its simplest form, a DNA sequence record can be represented as a string of nucleotides with some tag or identifier. Here is a nucleotide file as represented in a FASTA (or Pearson format) file:

```
>L04459
GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTACC
GCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGCG
ACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTT
ATTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCA
GCCAGTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTT
ATTTAAAGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGG
CCTATACACATAGACATTTGCACCTTATACATATAC
```

Similarly, for a protein record, the FASTA record would appear as follows:

```
>P31373
MTLQESDKFATKAIHAGEHVDVHGSVIEPISLSTTFKQSSPANPIGTYEYSRSQNPNRENLERAVAALENAQ
YGLAFSSGSATTATILQSLPQGSHAVSIGDVYGGTHRYFTKVANAHGVETSFTNDLLNDLPQLIKENTKLVW
IETPTNPTLKVTDIQKVADLIKKHAAGQDVILVVDNTFLSPYISNPLNFGADIVVHSATKYINGHSDVVLGV
LATNNKPLYERLQFLQNAIGAIPSPFDAWLTHRGLKTLHLRVRQAALSANKIAEFLAADKENVVAVNYPGLK
THPNYDVVLKQHRDALGGGMISFRIKGGAEAASKFASSTRLFTLAESLGGIESLLEVPAVMTHGGIPKEARA
SGVFDDLVRISVGIEDTDDLLEDIKQALKQATN
```

The FASTA format is used in a variety of molecular biology software suites. In its simplest incarnation (as shown above) the ''greater than'' character (>) designates the beginning of a new file. An identifier (L04459 in the first of the preceding examples) is followed by the DNA sequence in lowercase or uppercase letters, usually with 60 characters per line. Users and databases can then, if they wish, add a certain degree of complexity to this format. For example, without breaking any of the rules just outlined, one could add more information to the FASTA definition line, making the simple format a little more informative, as follows:

```
>gi|171361|gb|L04459|YSCCYS3A Saccharomyces cerevisiae cystathionine gamma-lyase
(CYS3) gene, complete cds.
GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTACC
GCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGCG
ACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTT
ATTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCA
GCCAGTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTT
ATTTAAAGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGG
CCTATACACATAGACATTTGCACCTTATACATATAC
```

This modified FASTA file now has the gi number (see below and Chapter 2), the GenBank accession number, the LOCUS name, and the DEFINITION line from the GenBank record. The record was passed from the underlying ASN.1 record (see Appendix 3.2), which NCBI uses to actually store and maintain all its data.

Over the years, many file formats have come and gone. Tools exist to convert the sequence itself into the minimalist view of one format or another. NCBI's asn2ff (ASN.1 to flatfile) will convert an ASN.1 file into a variety of flatfiles. The asn2ff program will generate GenBank, EMBL, GenPept, SWISS-PROT, and FASTA formats and is available from the NCBI Toolkit. READSEQ is another tool that has been widely used and incorporated into many work environments. Users should be aware that the features from a GenBank or EMBL format may be lost when passed through such utilities. Programs that need only the sequence (e.g., BLAST; see Chapter 8) are best used with a FASTA format for the query sequence. Although less informative than other formats, the FASTA format offers a simple way of dealing with the primary data in a human- and computer-readable fashion.

THE DATABASE

A full release of GenBank occurs on a bimonthly schedule with incremental (and nonincremental) daily updates available by anonymous FTP. The International Nucleotide Sequence Database Collaboration also exchanges new and updated records daily. Therefore, all sequences present in GenBank are also present in DDBJ and EMBL, as described in the introduction to this chapter. The three databases rely on a common data format for information described in the feature table documentation (see below). This represents the *lingua franca* for nucleotide sequence database annotations. Together, the nucleotide sequence databases have developed defined submission procedures (see Chapter 4), a series of guidelines for the content and format of all records.

As mentioned above, nucleotide records are often the primary source of sequence and biological information from which protein sequences in the protein databases are derived. There are three important consequences of not having the correct or proper information on the nucleotide record:

- If a coding sequence is not indicated on a nucleic acid record, it will not be represented in the protein databases. Thus, because querying the protein databases is the most sensitive way of doing similarity discoveries (Chapter 8), failure to indicate the CDS intervals on an mRNA or genomic sequence of interest (when one should be present) may cause important discoveries to be missed.
- The set of features usable in the nucleotide feature table that are specific to protein sequences themselves is limited. Important information about the protein will not be entered in the records in a ''parsable place.'' (The information may be present in a note, but it cannot reliably be found in the same place under all circumstances.)
- If a coding feature on a nucleotide record contains incorrect information about the protein, this could be propagated to other records in both the nucleotide and protein databases on the basis of sequence similarity.

THE GENBANK FLATFILE: A DISSECTION

The GenBank flatfile (GBFF) is the elementary unit of information in the GenBank database. It is one of the most commonly used formats in the representation of

biological sequences. At the time of this writing, it is the format of exchange from GenBank to the DDBJ and EMBL databases and vice versa. The DDBJ flatfile format and the GBFF format are now nearly identical to the GenBank format (Appendix 3.1). Subtle differences exist in the formatting of the definition line and the use of the gene feature. EMBL uses line-type prefixes, which indicate the type of information present in each line of the record (Appendix 3.2). The feature section (see below), prefixed with `FT`, is identical in content to the other databases. All these formats are really reports from what is represented in a much more structured way in the underlying ASN.1 file.

The GBFF can be separated into three parts: the *header*, which contains the information (descriptors) that apply to the whole record; the *features*, which are the annotations on the record; and the nucleotide sequence itself. All major nucleotide database flat files end with // on the last line of the record.

The Header

The header is the most database-specific part of the record. The various databases are not obliged to carry the same information in this segment, and minor variations exist, but some effort is made to ensure that the same information is carried from one to the other. The first line of all GBFFs is the Locus line:

```
LOCUS       AF111785     5925 bp    mRNA            PRI       01-SEP-1999
```

The first element on this line is the locus name. This element was historically used to represent the locus that was the subject of the record, and submitters and database staff spent considerable time in devising it so that it would serve as a mnemonic. Characters after the first can be numerical or alphabetic, and all letters are uppercase. The locusname was most useful back when most DNA sequence records represented only one genetic locus, and it was simple to find in GenBank a unique name that could represent the biology of the organism in a few letters and numbers. Classic examples include HUMHBB for the human β-globin locus or SV40 for the Simian virus (one of the copies anyway; there are many now). To be usable, the locus name needs to be unique within the database; because virtually all the meaningful designators have been taken, the LOCUS name has passed its time as a useful format element. Nowadays, this element must begin with a letter, and its length cannot exceed 10 characters. Because so many software packages rely on the presence of a unique LOCUS name, the databases have been reluctant to remove it altogether. The preferred path has been to instead put a unique word, and the simplest way to do this has been to use an accession number of ensured uniqueness: AF111785 in the example above conforms to the locus name requirement.

The second item on the locus line is the length of the sequence. Sequences can range from 1 to 350,000 base pairs (bp) in a single record. In practice, GenBank and the other databases seldom accept sequences shorter than 50 bp; therefore, the inclusion of polymerase chain reaction (PCR) primers as sequences (i.e., submissions of 24 bp) is discouraged. The 350 kb limit is a practical one, and the various databases represent longer contigs in a variety of different and inventive ways (see Chapters 2 and 6 and Appendix 3.3). Records of greater than 350 kb are acceptable in the database if the sequence represents a single gene.

The third item on the locus line indicates the molecule type. The ''mol type'' usually is DNA or RNA, and it can also indicate the strandedness (single or double, as ss or ds, respectively); however, these attributes are rarely used these days (another historical leftover). The acceptable mol types are DNA, RNA, tRNA, rRNA, mRNA, and uRNA and are intended to represent the original biological molecule. For example, a cDNA that is sequenced really represents an mRNA, and mRNA is the indicated mol type for such a sequence. If the tRNA or rRNA has been sequenced directly or via some cDNA intermediate, then tRNA or rRNA is shown as the mol type. If the ribosomal RNA gene sequence was obtained via the PCR from genomic DNA, then DNA is the mol type, even if the sequence encodes a structural RNA.

The fourth item on the locus line is the GenBank division code: three letters, which have either taxonomic inferences or other classification purposes. Again, these codes exist for historical reasons, recalling the time when the various GenBank divisions were used to break up the database files into what was then a more manageable size. The GenBank divisions are slightly different from those of EMBL or DDBJ, as described elsewhere (Ouellette and Boguski, 1997). NCBI has not introduced additional organism-based divisions in quite a few years, but new, function-based divisions have been very useful because they represent functional and definable sequence types (Ouellette and Boguski, 1997). The Expressed Sequence Tags (EST) division was introduced in 1993 (Boguski et al., 1993) and was soon followed by a division for Sequence Tagged Sites (STS). These, along with the Genome Survey Sequences (GSS) and unfinished, High Throughput Genome sequences (HTG), represent functional categories that need to be dealt with by the users and the database staff in very different ways. For example, a user can query these data sets specifically (e.g., via a BLASTN search against the EST or HTG division). Knowing that the hit is derived from a specific technique-oriented database allows one to interpret the data accordingly. At this time, GenBank, EMBL, and DDBJ interpret the various functional divisions in the same way, and all data sets are represented in the same division from one database to the next. The CON division is a new division for constructed (or ''contigged'') records. This division contains segmented sets as well as all large assemblies, which may exceed (sometimes quite substantially) the 350,000-bp limit presently imposed on single records. Such records may take the form shown in Appendix 3.3. The record from the CON division shown in Appendix 3.3 gives the complete genomic sequence of *Mycoplasma pneumoniae*, which is more than 800,000 base pairs in length. This CON record does not include sequences or annotations; rather, it includes instructions on how to assemble pieces present in other divisions into larger or assembled pieces. Records within the CON division have accession and version numbers and are exchanged, like all other records within the collaboration.

The date on the locus line is the date the record was last made public. If the record has not been updated since being made public, the date would be the date that it was first made public. If any of the features or annotations were updated and the record was rereleased, then the date corresponds to the last date the entry was released. Another date contained in the record is the date the record was submitted (see below) to the database. It should be noted that none of these dates is legally binding on the promulgating organization. The databases make no claim that the dates are error-free; they are included as guides to users and should not be submitted in any arbitration dispute. To the authors' knowledge, they have never been used in establishing priority and publication dates for patent application.

```
DEFINITION Homo sapiens myosin heavy chain IIx/d mRNA, complete cds.
```

The definition line (also referred to as the ''def line'') is the line in the GenBank record that attempts to summarize the biology of the record. This is the line that appears in the FASTA files that NCBI generates and is what is seen in the summary line for BLAST hits generated from a BLAST similarity search (Chapter 8). Much care is taken in the generation of these lines, and, although many of them can be generated automatically from the features in the record, they are still reviewed by the database staff to make sure that consistency and richness of information are maintained. Nonetheless, it is not always possible to capture all the biology in a single line of text, and databases cope with this in a variety of ways. There are some agreements in force between the databases, and the databases are aware of each other's guidelines and try to conform to them.

The generalized syntax for an mRNA definition line is as follows:

```
Genus species product name (gene symbol) mRNA, complete cds.
```

The generalized syntax for a genomic record is

```
Genus species product name (gene symbol) gene, complete cds.
```

Of course, records of many other types of data are accounted for by the guidelines used by the various databases. The following set of rules, however, applies to organelle sequences, and these rules are used to ensure that the biology and source of the DNA are clear to the user and to the database staff (assuming they are clear to the submitter):

```
DEFINITION Genus species protein X(xxx) gene, complete cds;
    [one choice from below], OR
DEFINITION Genus species XXS ribosomal RNA gene, complete sequence;
    [one choice from below].
  nuclear gene(s) for mitochondrial product(s)
  nuclear gene(s) for chloroplast product(s)
  mitochondrial gene(s) for mitochondrial product(s)
  chloroplast gene(s) for chloroplast product(s)
```

In accordance with a recent agreement among the collaborative databases, the full genus-species names are given in the definition lines; common names (e.g., human) or abbreviated genus names (e.g., *H. sapiens* for Homo sapiens) are no longer used. The many records in the database that precede this agreement will eventually be updated. One organism has escaped this agreement: the human immunodeficiency virus is to be represented in the definition line as HIV1 and HIV2.

```
ACCESSION   AF111785
```

The accession number, on the third line of the record, represents the primary key to reference a given record in the database. This is the number that is cited in publications and is always associated with this record; that is, if the sequence is updated (e.g., by changing a single nucleotide), the accession number will not change. At this time, accession numbers exist in one of two formats: the ''1 + 5''

and "2 + 6" varieties, where 1 + 5 indicates one uppercase letter followed by five digits and 2 + 6 is two letters plus six digits. Most of the new records now entering the databases are of the latter variety. All GenBank records have only a single line with the word ACCESSION on it; however, there may be more than one accession number. The vast majority of records only have one accession number. This number is always referred to as the primary accession number; all others are secondary. In cases where more than one accession number is shown, the first accession number is the primary one.

Unfortunately, secondary accession numbers have meant a variety of things over the years, and no single definition applies. The secondary accession number may be related to the primary one, or the primary accession number may be a replacement for the secondary, which no longer exists. There is an ongoing effort within the Collaboration to make the latter the default for all cases, but, because secondary accession numbers have been used for more than 15 years (a period during which the management of GenBank changed), all data needed to elucidate all cases are not available.

```
ACCESSION   AF111785
VERSION     AF111785.1 GI:4808814
```

The version line contains the Accession.version and the gi (geninfo identifier). These identifiers are associated with a unique nucleotide sequence. Protein sequences also have accession numbers (protein_ids). These are also represented as Accession.version and gi numbers for unique sequences (see below). If the sequence changes, the version number in the Accession.version will be incremented by one and the gi will change (although not by one, but to the next available integer). The accession number stays the same. The example above shows version 1 of the sequence having accession number AF111785 and gi number 4808814.

KEYWORDS

The keywords line is another historical relic that is, in many cases, unfortunately misused. Adding keywords to an entry is often not very useful because over the years so many authors have selected words not on a list of controlled vocabulary and not uniformly applied to the whole database. NCBI, therefore, discourages the use of keywords but will include them on request, especially if the words are not present elsewhere in the record or are used in a controlled fashion (e.g., for EST, STS, GSS, and HTG records). At this time, the resistance to adding keywords is a matter of policy at NCBI/GenBank only.

```
SOURCE      human.
ORGANISM    Homo sapiens
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
 Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
```

The source line will either have the common name for the organism or its scientific name. Older records may contain other source information (see below) in this field. A concerted effort is now under way to assure that all other information present in the source feature (as opposed to the source line) and all lines in the taxonomy block (source and organism lines) can be derived from what is in the source feature

and the taxonomy server at NCBI. Those interested in the lineage and other aspects of the taxonomy are encouraged to visit the taxonomy home page at NCBI. This taxonomy database is used by all nucleotide sequence databases, as well as SWISS-PROT.

```
REFERENCE 1 (bases 1 to 5925)
AUTHORS   Weiss,A., McDonough,D., Wertman,B., Acakpo-Satchivi,L.,
          Montgomery,K., Kucherlapati,R., Leinwand,L. and Krauter,K.
TITLE     Organization of human and mouse skeletal myosin heavy chain
          gene clusters is highly conserved
JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 96 (6), 2958-2963 (1999)
MEDLINE   99178997
PUBMED    10077619
```

Each GenBank record must have at least one reference or citation. It offers scientific credit and sets a context explaining why this particular sequence was determined. In many cases, the record will have two or more reference blocks, as shown in Appendix 3.1. The preceding sample indicates a published paper. There is a MEDLINE and PubMed identifier present that provides a link to the MEDLINE/PubMed databases (see Chapter 7). Other references may be annotated as unpublished (which could be ''submitted) or as placeholders for a publication, as shown.

```
REFERENCE 1 (bases 1 to 3291)
AUTHORS   Morcillo, P., Rosen, C.,Baylies, M.K. and Dorsett, D.
TITLE     CHIP, a widely expressed chromosomal protein required for
          remote enhancer activity and segmentation in Drosophila
JOURNAL   Unpublished
REFERENCE 3 (bases 1 to 5925)
AUTHORS   Weiss,A. and Leinwand,L.A.
TITLE     Direct Submission
JOURNAL   Submitted (09-DEC-1998) MCDB, University of Colorado at
          Boulder, Campus Box 0347, Boulder, Colorado 80309-0347, USA
```

The last citation is present on most GenBank records and gives scientific credit to the people responsible for the work surrounding the submitted sequence. It usually includes the postal address of the first author or the lab where the work was done. The date represents the date the record was submitted to the database but not the date on which the data were first made public, which is the date on the locus line if the record was not updated. Additional submitter blocks may be added to the record each time the sequences are updated.

The last part of the header section in the GBFF is the comment. This section includes a great variety of notes and comments (also called ''descriptors'') that refer to the whole record. Genome centers like to include their contact information in this section as well as give acknowledgments. This section is optional and not found in most records in GenBank. The comment section may also include E-mail addresses or URLs, but this practice is discouraged at NCBI (although certain exceptions have been made for genome centers as mentioned above). The simple reason is that E-mail addresses tend to change more than the postal addresses of buildings. DDBJ has been including E-mail addresses for some years, again representing a subtle difference in policy. The comment section also contains information about the history

of the sequence. If the sequence of a particular record is updated, the comment will contain a pointer to the previous version of the record.

```
COMMENT On Dec 23, 1999 this sequence version replaced gi:4454562.
```

Alternatively, if you retrieve an earlier version of the record, this comment will point forward to the newer version of the sequence and also backward if there was an earlier still version

```
COMMENT [WARNING] On Dec 23, 1999 this sequence was replaced by
        a newer version gi:6633795.
```

The Feature Table

The middle segment of the GBFF record, the feature table, is the most important direct representation of the biological information in the record. One could argue that the biology is best represented in the bibliographic reference, cited by the record. Nonetheless, a full set of annotations within the record facilitates quick extraction of the relevant biological features and allows the submitter to indicate why this record was submitted to the database. What becomes relevant here is the choice of annotations presented in this section. The GenBank feature table documentation describes in great detail the legal features (i.e., the ones that are allowed) and what qualifiers are permitted with them. This, unfortunately, has often invited an excess of invalid, speculative, or computed annotations. If an annotation is simply computed, its usefulness as a comment within the record is diminished.

Described below are some of the key GenBank features, with information on why they are important and what information can be extracted from them. The discussion here is limited to the biological underlyings of these features and guidelines applied to this segment by the NCBI staff. This material will also give the reader some insight into the NCBI data model (Chapter 2) and the important place the GBFF occupies in the analysis of sequences, serving also to introduce the concept of features and qualifiers in GenBank language. The features are slightly different from other features discussed in Chapter 2. In the GBFF report format, any component of this section designated as ''feature.'' In the NCBI data model, ''features'' refer to annotations that are on a part of the sequences, whereas annotations that describe the whole sequence are called ''descriptors.'' Thus, the source feature in the GenBank flatfile is really a descriptor in the data model view (the BioSource, which refers to the whole sequence), not a feature as used elsewhere. Because this is a chapter on the GenBank database, the ''feature'' will refer to all components of the feature table. The readers should be aware of this subtle difference, especially when referring to other parts of this book.

The Source Feature. The source feature is the only feature that must be present on all GenBank records. All features have a series of legal qualifiers, some of which are mandatory (e.g., `/organism` for source). All DNA sequence records have some origin, even if synthetic in the extreme case. In most cases, there will be a single source feature, and it will contain the `/organism`. Here is what we have in the example from Appendix 3.1:

```
source 1..5925
       /organism="Homo sapiens"
       /db_xref="taxon:9606"
       /chromosome="17"
       /map="17p13.1"
       /tissue_type="skeletal muscle"
```

The organism qualifier contains the scientific genus and species name. In some cases, ''organisms'' can be described at the subspecies level. For the source feature, the series of qualifiers will contain all matters relating to the BioSource, and these may include mapping, chromosome or tissue from which the molecule that was sequenced was obtained, clone identification, and other library information. For the source feature, as is true for all features in a GenBank record, care should be taken to avoid adding superfluous information to the record. For the reader of these records, anything that cannot be computationally validated should be taken with a grain of salt. Tissue source and library origin are only as good as the controls present in the associated publication (if any such publication exists) and only insofar as that type of information is applied uniformly across all records in GenBank. With sets of records in which the qualifiers are applied in a systematic way, as they are for many large EST sets, the taxonomy can be validated (i.e., the organism does exist in the database of all organisms that is maintained at the NCBI). If, in addition, the qualifier is applied uniformly across all records, it is of value to the researcher. Unfortunately, however, many qualifiers are derived without sufficient uniformity across the database and hence are of less value.

Implicit in the BioSource and the organism that is assigned to it is the genetic code used by the DNA/RNA, which will be used to translate the nucleic acid to represent the protein sequence (if one is present in the record). This information is shown on the CDS feature.

The CDS Feature. The CDS feature contains instructions to the reader on how to join two sequences together or on how to make an amino acid sequence from the indicated coordinates and the inferred genetic code. The GBFF view, being as DNA-centric as it is, maps all features through a DNA sequence coordinate system, not that of amino acid reference points, as in the following example from GenBank accession X59698 (contributed by a submission to EMBL).

```
sig peptide 160..231
CDS         160..>2301
            /codon_start=1
            /product="EGF-receptor"
            /protein_id="CAA42219.1"
            /db_xref="GI:50804"
            /db_xref="MGD:MGI:95294"
            /db_xref="SWISS-PROT:Q01279"
            /translation="MRPSGTARTTLLVLLTALCAAGGALEEKKVCQGTSNRLTQLGTF
            EDHFLSLQRMYNNCEVVLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLE
            NLQIIRGNALYENTYALAILSNYGTNRTGLRELPMRNLQEILIGAVRFSNNPILCNMD
            TIQWRDIVQNVFMSNMSMDLQSHPSSCPKCDPSCPNGSCWGGGEENCQKLTKIICAQQ
            CSHRCRGRSPSDCCHNQCAAGCTGPRESDCLVCQKFQDEATCKDTCPPLMLYNPTTYQ
            MDVNPEGKYSFGATCVKKCPRNYVVTDHGSCVRACGPDYYEVEEDGIRKCKKCDGPCR
```

```
            KVCNGIGIGEFKDTLSINATNIKHFKYCTAISGDLHILPVAFKGDSFTRTPPLDPREL
            EILKTVKEITGFLLIQAWPDNWTDLHAFENLEIIRGRTKQHGQFSLAVVGLNITSLGL
            RSLKEISDGDVIISGNRNLCYANTINWKKLFGTPNQKTKIMNNRAEKDCKAVNHVCNP
            LCSSEGCWGPEPRDCVSCQNVSRGRECVEKWNILEGEPREFVENSECIQCHPECLPQA
            MNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGIMGENNTLVWKYADANNVCHLCHANC
            TYGCAGPGLQGCEVWPSGPKIPSIATGIVGGLLFIVVVALGIGLFMRRRHIVRKRTLR
            RLLQERELVEPLTPSGEAPNQAHLRILKETEF"
mat peptide 232..>2301
            /product="EGF-receptor"
```

This example also illustrates the use of the database cross-reference (`db_xref`). This controlled qualifier allows the databases to cross-reference the sequence in question to an external database (the first identifier) with an identifier used in that database. The list of allowed `db_xref` databases is maintained by the International Nucleotide Sequence Database Collaboration.

```
/protein_id="CAA42219.1"
/db_xref="GI:50804"
```

As mentioned above, NCBI assigns an accession number and a gi (geninfo) identifier to all sequences. This means that translation products, which are sequences in their own right (not simply attachments to a DNA record, as they are shown in a GenBank record), also get an accession number (`/protein_id`) and a gi number. These unique identifiers will change when the sequence changes. Each protein sequence is assigned a protein_id or protein accession number. The format of this accession number is ''3 + 5,'' or three letters and five digits. Like the nucleotide sequence accession number, the protein accession number is represented as Accession.version. The protein gi numbers appear as a gi `db_xref`. When the protein sequence in the record changes, the version of the accession number is incremented by one and the gi is also changed.

Thus, the version number of the accession number presents the user with an easy way to look up the previous version of the record, if one is present. Because amino acid sequences represent one of the most important by-products of the nucleotide sequence database, much attention is devoted to making sure they are valid. (If a translation is present in a GenBank record, there are valid coordinates present that can direct the translation of nucleotide sequence.) These sequences are the starting material for the protein databases and offer the most sensitive way of making new gene discoveries (Chapter 8). Because these annotations can be validated, they have added value, and having the correct identifiers also becomes important. The correct product name, or protein name, can be subjective and often is assigned via weak similarities to other poorly annotated sequences, which themselves have poor annotations. Thus, users should be aware of potential circular amplification of paucity of information. A good rule is that more information is usually obtained from records describing single genes or full-length mRNA sequences with which a published paper is associated. These records usually describe the work from a group that has studied a gene of interest in some detail. Fortunately, quite a few records of these types are in the database, representing a foundation of knowledge used by many.

The Gene Feature. The gene feature, which has been explicitly included in the GenBank flatfile for only a few years, has nevertheless been implicitly in use

since the beginning of the databases as a gene qualifier on a number of other features. By making this a separate feature, the de facto status has been made explicit, greatly facilitating the generation and validation of components now annotated with this feature. The new feature has also clearly shown in its short existence that biologists have very different definitions and uses for the gene feature in GenBank records. Although it is obvious that not all biologists will agree on a single definition of the gene feature, at its simplest interpretation, the gene feature represents a segment of DNA that can be identified with a name (e.g., the MyHC gene example from Appendix 3.1) or some arbitrary number, as is often used in genome sequencing project (e.g., T23J18.1 from GenBank accession number AC011661). The gene feature allows the user to see the gene area of interest and in some cases to select it.

The RNA Features. The various structural RNA features can be used to annotate RNA on genomic sequences (e.g., mRNA, rRNA, tRNA). Although these are presently not instantiated into separate records as protein sequences are, these sequences (especially the mRNA) are essential to our understanding of how higher genomes are organized. RNAs deserves special mention because they represent biological entities that can be measured in the lab and thus are pieces of information of great value for a genomic record and are often mRNA records on their own. This is in contrast to the promoter feature, which is poorly characterized, unevenly assigned in a great number of records, poorly defined from a biology point of view, and of lesser use in a GenBank record. The RNA feature on a genomic record should represent the experimental evidence of the presence of that biological molecule.

CONCLUDING REMARKS

The DDBJ/EMBL/GenBank database is the most commonly used nucleotide and protein sequence database. It represents a public repository of molecular biology information. Knowing what the various fields mean and how much biology can be obtained from these records greatly advances our understanding of this file format. Although the database was never meant to be read from computers, an army of computer-happy biologists have nevertheless parsed, converted, and extracted these records by means of entire suites of programs. THE DDBJ/EMBL/GenBank flatfile remains the format of exchange between the International Nucleotide Sequence Database Collaboration members, and this is unlikely to change for years to come, *despite* the availability of better, richer alternatives, such as the data described in ASN.1. However, therein lays the usefulness of the present arrangement: it is a readily available, simple format which can represent some abstraction of the biology it wishes to depict.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 3

GenBank Release Notes	*ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt*
READSEQ Sequence Conversion Tool	*http://magpie.bio.indiana.edu/MolecularBiology/Molbio_archive/readseq/*
Taxonomy Browser	*http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html*

TREMBL and Swiss-Prot Release Notes	*http://www.ebi.ac.uk/ebi_docs/swissprot_db/documentation.html*

REFERENCES

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucl. Acids Res.* 28, 45–48.

Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., and Tuli, M. A. (2000). The EMBL Nucleotide Sequence Database. *Nucl. Acids Res*. 28, 19–23.

Barker, W. C., Garavelli, J. S., Huang, H., McGarvey, P. B., Orcutt, B. C., Srinivasarao, G. Y., Xiao, C., Yeh, L. S., Ledley, R. S., Janda, J. F., Pfeiffer, F., Mewes, H.-W., Tsugita, A., and Wu, C. (2000). The Protein Information Resource (PIR). *Nucl. Acids Res.* 28, 41–44.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. and Wheeler, D. L. (1997). GenBank. *Nucl. Acids Res.* 25, 1–6.

Boguski, M. S., Lowe, T. M., Tolstoshev, C. M. (1993). dbEST—database for ''expressed sequence tags.'' *Nat. Genetics* 4: 332–333.

Cook-Deagan, R. (1993). *The Gene Wars. Science, Politics and the HumanGenome* (New York and London: W. W. Norton & Company).

Dayhoff, M. O., Eck, R. V., Chang, M. A., Sochard, M. R. (1965). *Atlas of Protein Sequence and Structure*. (National Biomedical Research Foundation, Silver Spring MD).

Mewes, H. W., Frischman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S., and Weil, B. (2000). MIPS: A database for genomes and protein sequences. *Nucl. Acids Res.* 28, 37–40.

Ouellette, B. F. F., and Boguski, M. S. (1997). Database divisions and homology search files: a guide for the perplexed. *Genome Res.* 7, 952–955.

Schuler, G. D., Epstein, J. A., Ohkawa, H., Kans, J. A. (1996). Entrez: Molecular biology database and retrieval system. *Methods Enzymol.* 266, 141–162.

Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H., and Gojobori, T. (1997). DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucl. Acids Res.* 28, 24–26.

APPENDICES

Appendix 3.1. Example of GenBank Flatfile Format

```
LOCUS       AF111785      5925 bp     mRNA            PRI       01-SEP-1999
DEFINITION  Homo sapiens myosin heavy chain IIx/d mRNA, complete cds.
ACCESSION   AF111785
VERSION     AF111785.1 GI:4808814
KEYWORDS    .
SOURCE      human.
ORGANISM    Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleos-
            tomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1 (bases 1 to 5925)
AUTHORS     Weiss,A., McDonough,D., Wertman,B., Acakpo-Satchivi,L., Mont-
            gomery,K., Kucherlapati,R., Leinwand,L. and Krauter,K.
TITLE       Organization of human and mouse skeletal myosin heavy chain gene
            clusters is highly conserved
JOURNAL     Proc. Natl. Acad. Sci. U.S.A. 96 (6), 2958-2963 (1999)
MEDLINE     99178997
PUBMED      10077619
```

```
REFERENCE   2 (bases 1 to 5925)
AUTHORS     Weiss,A., Schiaffino,S. and Leinwand,L.A.
TITLE       Comparative sequence analysis of the complete human sarcomeric
            myosin heavy chain family: implications for functional diversity
JOURNAL     J. Mol. Biol. 290 (1), 61-75 (1999)
MEDLINE     99318869
PUBMED      10388558
REFERENCE   3 (bases 1 to 5925)
AUTHORS     Weiss,A. and Leinwand,L.A.
TITLE       Direct Submission
JOURNAL     Submitted (09-DEC-1998) MCDB, University of Colorado at Boulder,
            Campus Box 0347, Boulder, Colorado 80309-0347, USA
FEATURES    Location/Qualifiers
source      1..5925
            /organism="Homo sapiens"
            /db_xref="taxon:9606"
            /chromosome="17"
            /map="17p13.1"
            /tissue_type="skeletal muscle"
CDS         1..5820
            /note="MyHC"
            /codon_start=1
            /product="myosin heavy chain IIx/d"
            /protein_id="AAD29951.1"
            /db_xref="GI:4808815"
            /translation="MSSDSEMAIFGEAAPFLRKSERERIEAQNKPFDAKTSVFVVDPK
            ESFVKATVQSREGGKVTAKTEAGATVTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAV
            LYNLKERYAAWMIYTYSGLFCVTVNPYKWLPVYNAEVVTAYRGKKRQEAPPHIFSISD
            NAYQFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAVTGEKKKEEVTSGKMQGT
            LEDQIISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLEKSRV
            TFQLKAERSYHIFYQIMSNKKPDLIEMLLITTNPYDYAFVSQGEITVPSIDDQEELMA
            TDSAIEILGFTSDERVSIYKLTGAVMHYGNMKFKQKQREEQAEPDGTEVADKAAYLQN
            LNSADLLKALCYPRVKVGNEYVTKGQTVQQVYNAVGALAKAVYDKMFLWMVTRINQQL
            DTKQPRQYFIGVLDIAGFEIFDFNSLEQLCINFTNEKLQQFFNHHMFVLEQEEYKKEG
            IEWTFIDFGMDLAACIELIEKPMGIFSILEEECMFPKATDTSFKNKLYEQHLGKSNNF
            QKPKPAKGKPEAHFSLIHYAGTVDYNIAGWLDKNKDPLNETVVGLYQKSAMKTLALLF
            VGATGAEAEAGGGKKGGKKKGSSFQTVSALFRENLNKLMTNLRSTHPHFVRCIIPNET
            KTPGAMEHELVLHQLRCNGVLEGIRICRKGFPSRILYADFKQRYKVLNASAIPEGQFI
            DSKKASEKLLGSIDIDHTQYKFGHTKVFFKAGLLGLLEEMRDEKLAQLITRTQAMCRG
            FLARVEYQKMVERRESIFCIQYNVRAFMNVKHWPWMKLYFKIKPLLKSAETEKEMANM
            KEEFEKTKEELAKTEAKRKELEEKMVTLMQEKNDLQLQVQAEADSLADAEERCDQLIK
            TKIQLEAKIKEVTERAEDEEEINAELTAKKRKLEDECSELKKDIDDLELTLAKVEKEK
            HATENKVKNLTEEMAGLDETIAKLTKEKKALQEAHQQTLDDLQAEEDKVNTLTKAKIK
            LEQQVDDLEGSLEQEKKIRMDLERAKRKLEGDLKLAQESAMDIENDKQQLDEKLKKKE
            FEMSGLQSKIEDEQALGMQLQKKIKELQARIEELEEEIEAERASRAKAEKQRSDLSRE
            LEEISERLEEAGGATSAQIEMNKKREAEFQKMRRDLEEATLQHEATAATLRKKHADSV
            AELGEQIDNLQRVKQKLEKEKSEMKMEIDDLASNMETVSKAKGNLEKMCRALEDQLSE
            IKTKEEEQQRLINDLTAQRARLQTESGEYSRQLDEKDTLVSQLSRGKQAFTQQIEELK
            RQLEEEIKAKSALAHALQSSRHDCDLLREQYEEEQEAKAELQRAMSKANSEVAQWRTK
            YETDAIQRTEELEEAKKKLAQRLQDAEEHVEAVNAKCASLEKTKQRLQNEVEDLMIDV
            ERTNAACAALDKKQRNFDKILAEWKQKCEETHAELEASQKESRSLSTELFKIKNAYEE
            SLDQLETLKRENKNLQQEISDLTEQIAEGGKRIHELEKIKKQVEQEKSELQAALEEAE
            ASLEHEEGKILRIQLELNQVKSEVDRKIAEKDEEIDQMKRNHIRIVESMQSTLDAEIR
            SRNDAIRLKKKMEGDLNEMEIQLNHANRMAAEALRNYRNTQAILKDTQLHLDDALRSQ
            EDLKEQLAMVERRANLLQAEIEELRATLEQTERSRKIAEQELLDASERVQLLHTQNTS
            LINTKKKLETDISQIQGEMEDIIQEARNAEEKAKKAITDAAMMAEELKKEQDTSAHLE
            RMKKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEGEVESEQKRNVEAVK
            GLRKHERKVKELTYQTEEDRKNILRLQDLVDKLQAKVKSYKRQAEEAEEQSNVNLSKF
            RRIQHELEEAEERADIAESQVNKLRVKSREVHTKIISEE"
```

```
BASE COUNT     1890 a 1300 c 1613 g 1122 t
ORIGIN
     1 atgagttctg actctgagat ggccattttt ggggaggctg ctcctttcct ccgaaagtct
    61 gaaagggagc gaattgaagc ccagaacaag ccttttgatg ccaagacatc agtctttgtg
   121 gtggacccta aggagtcctt tgtgaaagca acagtgcaga gcagggaagg ggggaaggtg
<< Sequence deleted to save space >>
  5701 cggaggatcc agcacgagct ggaggaggcc gaggaaaggg ctgacattgc tgagtcccag
  5761 gtcaacaagc tgagggtgaa gagcagggag gttcacacaa aaatcataag tgaagagtaa
  5821 tttatctaac tgctgaaagg tgaccaaaga aatgcacaaa atgtgaaaat ctttgtcact
  5881 ccattttgta cttatgactt ttggagataa aaaatttatc tgcca
//
```

Appendix 3.2. Example of EMBL Flatfile Format

```
ID AF111785 standard; RNA; HUM; 5925 BP.
XX
AC AF111785;
XX
SV AF111785.1
XX
DT 13-MAY-1999 (Rel. 59, Created)
DT 07-SEP-1999 (Rel. 61, Last updated, Version 3)
XX
DE Homo sapiens myosin heavy chain IIx/d mRNA, complete cds.
XX
KW .
XX
OS Homo sapiens (human)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN [1]
RP 1-5925
RX MEDLINE; 99178997.
RA Weiss A., McDonough D., Wertman B., Acakpo-Satchivi L., Montgomery K.,
RA Kucherlapati R., Leinwand L., Krauter K.;
RT "Organization of human and mouse skeletal myosin heavy chain gene clusters
RT is highly conserved";
RL Proc. Natl. Acad. Sci. U.S.A. 96(6):2958-2963(1999).
XX
RN [2]
RP 1-5925
RX MEDLINE; 99318869.
RA Weiss A., Schiaffino S., Leinwand L.A.;
RT "Comparative sequence analysis of the complete human sarcomeric myosin
RT heavy chain family: implications for functional diversity";
RL J. Mol. Biol. 290(1):61-75(1999).
XX
RN [3]
RP 1-5925
RA Weiss A., Leinwand L.A.;
RT ;
RL Submitted (09-DEC-1998) to the EMBL/GenBank/DDBJ databases.
RL MCDB, University of Colorado at Boulder, Campus Box 0347, Boulder, Colorado
RL 80309-0347, USA
```

```
XX
DR SPTREMBL; Q9Y622; Q9Y622.
XX
FH Key Location/Qualifiers
FH
FT source    1..5925
FT           /chromosome="17"
FT           /db_xref="taxon:9606"
FT           /organism="Homo sapiens"
FT           /map="17p13.1"
FT           /tissue_type="skeletal muscle"
FT CDS       1..5820
FT           /codon_start=1
FT           /db_xref="SPTREMBL:Q9Y622"
FT           /note="MyHC"'
FT           /product="myosin heavy chain IIx/d"
FT           /protein_id="AAD29951.1"
FT           /translation="MSSDSEMAIFGEAAPFLRKSERERIEAQNKPFDAKTSVFVVDPKE
FT           SFVKATVQSREGGKVTAKTEAGATVTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLY
FT           NLKERYAAWMIYTYSGLFCVTVNPYKWLPVYNAEVVTAYRGKKRQEAPPHIFSISDNAY
FT           QFMLTDRENQSILITGESGAGKTVNTKRVIQYFATIAVTGEKKKEEVTSGKMQGTLEDQ
FT           IISANPLLEAFGNAKTVRNDNSSRFGKFIRIHFGTTGKLASADIETYLLEKSRVTFQLK
FT           AERSYHIFYQIMSNKKPDLIEMLLITTNPYDYAFVSQGEITVPSIDDQEELMATDSAIE
FT           ILGFTSDERVSIYKLTGAVMHYGNMKFKQKQREEQAEPDGTEVADKAAYLQNLNSADLL
FT           KALCYPRVKVGNEYVTKGQTVQQVYNAVGALAKAVYDKMFLWMVTRINQQLDTKQPRQY
FT           FIGVLDIAGFEIFDFNSLEQLCINFTNEKLQQFFNHHMFVLEQEEYKKEGIEWTFIDFG
FT           MDLAACIELIEKPMGIFSILEEECMFPKATDTSFKNKLYEQHLGKSNNFQKPKPAKGKP
FT           EAHFSLIHYAGTVDYNIAGWLDKNKDPLNETVVGLYQKSAMKTLALLFVGATGAEAEAG
FT           GGKKGGKKKGSSFQTVSALFRENLNKLMTNLRSTHPHFVRCIIPNETKTPGAMEHELVL
FT           HQLRCNGVLEGIRICRKGFPSRILYADFKQRYKVLNASAIPEGQFIDSKKASEKLLGSI
FT           DIDHTQYKFGHTKVFFKAGLLGLLEEMRDEKLAQLITRTQAMCRGFLARVEYQKMVERR
FT           ESIFCIQYNVRAFMNVKHWPWMKLYFKIKPLLKSAETEKEMANMKEEFEKTKEELAKTE
FT           AKRKELEEKMVTLMQEKNDLQLQVQAEADSLADAEERCDQLIKTKIQLEAKIKEVTERA
FT           EDEEEINAELTAKKRKLEDECSELKKDIDDLELTLAKVEKEKHATENKVKNLTEEMAGL
FT           DETIAKLTKEKKALQEAHQQTLDDLQAEEDKVNTLTKAKIKLEQQVDDLEGSLEQEKKI
FT           RMDLERAKRKLEGDLKLAQESAMDIENDKQQLDEKLKKKEFEMSGLQSKIEDEQALGMQ
FT           LQKKIKELQARIEELEEEIEAERASRAKAEKQRSDLSRELEEISERLEEAGGATSAQIE
FT           MNKKREAEFQKMRRDLEEATLQHEATAATLRKKHADSVAELGEQIDNLQRVKQKLEKEK
FT           SEMKMEIDDLASNMETVSKAKGNLEKMCRALEDQLSEIKTKEEEQQRLINDLTAQRARL
FT           QTESGEYSRQLDEKDTLVSQLSRGKQAFTQQIEELKRQLEEEIKAKSALAHALQSSRHD
FT           CDLLREQYEEEQEAKAELQRAMSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQRLQ
FT           DAEEHVEAVNAKCASLEKTKQRLQNEVEDLMIDVERTNAACAALDKKQRNFDKILAEWK
FT           QKCEETHAELEASQKESRSLSTELFKIKNAYEESLDQLETLKRENKNLQQEISDLTEQI
FT           AEGGKRIHELEKIKKQVEQEKSELQAALEEAEASLEHEEGKILRIQLELNQVKSEVDRK
FT           IAEKDEEIDQMKRNHIRIVESMQSTLDAEIRSRNDAIRLKKKMEGDLNEMEIQLNHANR
FT           MAAEALRNYRNTQAILKDTQLHLDDALRSQEDLKEQLAMVERRANLLQAEIEELRATLE
FT           QTERSRKIAEQELLDASERVQLLHTQNTSLINTKKKLETDISQIQGEMEDIIQEARNAE
FT           EKAKKAITDAAMMAEELKKEQDTSAHLERMKKNLEQTVKDLQHRLDEAEQLALKGGKKQ
FT           IQKLEARVRELEGEVESEQKRNVEAVKGLRKHERKVKELTYQTEEDRKNILRLQDLVDK
FT           LQAKVKSYKRQAEEAEEQSNVNLSKFRRIQHELEEAEERADIAESQVNKLRVKSREVHT
FT           KIISEE"
XX
SQ Sequence 5925 BP; 1890 A; 1300 C; 1613 G; 1122 T; 0 other;
 atgagttctg actctgagat ggccattttt ggggaggctg ctcctttcct ccgaaagtct 60
 gaaagggagc gaattgaagc ccagaacaag ccttttgatg ccaagacatc agtctttgtg 120
  << Sequence deleted to save space >>
 cggaggatcc agcacgagct ggaggaggcc gaggaaaggg ctgacattgc tgagtcccag 5760
 gtcaacaagc tgagggtgaa gagcagggag gttcacacaa aaatcataag tgaagagtaa 5820
```

```
 tttatctaac tgctgaaagg tgaccaaaga aatgcacaaa atgtgaaaat ctttgtcact 5880
 ccattttgta cttatgactt ttggagataa aaaatttatc tgcca 5925
//
```

Appendix 3.3. Example of a Record in CON Division

```
LOCUS        U00089      816394 bp     DNA    circular   CON         10-MAY-1999
DEFINITION   Mycoplasma pneumoniae M129 complete genome.
ACCESSION    U00089
VERSION      U00089.1 GI:6626256
KEYWORDS     .
SOURCE       Mycoplasma pneumoniae.
ORGANISM     Mycoplasma pneumoniae
             Bacteria; Firmicutes; Bacillus/Clostridium group; Mollicutes;
             Mycoplasmataceae; Mycoplasma.
REFERENCE    1 (bases 1 to 816394)
AUTHORS      Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and
             Herrmann,R.
TITLE        Complete sequence analysis of the genome of the bacterium Mycoplasma
             pneumoniae
JOURNAL      Nucleic Acids Res. 24 (22), 4420-4449 (1996)
MEDLINE      97105885
REFERENCE    2 (bases 1 to 816394)
AUTHORS      Himmelreich,R., Hilbert,H. and Li,B.-C.
TITLE        Direct Submission
JOURNAL      Submitted (15-NOV-1996) Zentrun fuer Molekulare Biologie Heidelberg,
             University Heidelberg, 69120 Heidelberg, Germany
FEATURES     Location/Qualifiers
source       1..816394
             /organism="Mycoplasma pneumoniae"
             /strain="M129"
             /db_xref="taxon:2104"
             /note="ATCC 29342"
CONTIG       join(AE000001.1:1..9255,AE000002.1:59..16876,AE000003.1:59..10078,
             AE000004.1:59..17393,AE000005.1:59..10859,AE000006.1:59..11441,
             AE000007.1:59..10275,AE000008.1:59..9752,AE000009.1:59..14075,
             AE000010.1:59..11203,AE000011.1:59..15501,AE000012.1:59..10228,
             AE000013.1:59..10328,AE000014.1:59..12581,AE000015.1:59..17518,
             AE000016.1:59..16518,AE000017.1:59..18813,AE000018.1:59..11147,
             AE000019.1:59..10270,AE000020.1:59..16613,AE000021.1:59..10701,
             AE000022.1:59..12807,AE000023.1:59..13289,AE000024.1:59..9989,
             AE000025.1:59..10770,AE000026.1:59..11104,AE000027.1:59..33190,
             AE000028.1:59..10560,AE000029.1:59..10640,AE000030.1:59..11802,
             AE000031.1:59..11081,AE000032.1:59..12622,AE000033.1:59..12491,
             AE000034.1:59..11844,AE000035.1:59..10167,AE000036.1:59..11865,
             AE000037.1:59..11391,AE000038.1:59..11399,AE000039.1:59..14233,
             AE000040.1:59..13130,AE000041.1:59..11259,AE000042.1:59..12490,
             AE000043.1:59..11643,AE000044.1:59..15473,AE000045.1:59..10855,
             AE000046.1:59..11562,AE000047.1:59..20217,AE000048.1:59..10109,
             AE000049.1:59..12787,AE000050.1:59..12516,AE000051.1:59..16249,
             AE000052.1:59..12390,AE000053.1:59..10305,AE000054.1:59..10348,
             AE000055.1:59..9893,AE000056.1:59..16213,AE000057.1:59..11119,
             AE000058.1:59..28530,AE000059.1:59..12377,AE000060.1:59..11670,
             AE000061.1:59..24316,AE000062.1:59..10077,AE000063.1:59..1793)
//
```

4

SUBMITTING DNA SEQUENCES TO THE DATABASES

Jonathan A. Kans

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland

B. F. Francis Ouellette

Centre for Molecular Medicine and Therapeutics
Children's and Women's Health Centre of British Columbia
University of British Columbia
Vancouver, British Columbia

INTRODUCTION

DNA sequence records from the public databases (DDBJ/EMBL/GenBank) are essential components of computational analysis in molecular biology. The sequence records are also reagents for improved curated resources like LocusLink (see Chapter 7) or many of the protein databases. Accurate and informative biological annotation of sequence records is critical in determining the function of a disease gene by sequence similarity search. The names or functions of the encoded protein products, the name of the genetic locus, and the link to the original publication of that sequence make a sequence record of immediate value to the scientist who retrieves it as the result of a BLAST or Entrez search. Effective interpretation of recently finished human genome sequence data is only possible by making use of *all* submitted data provided along with the actual sequence. These complete, annotated records capture the biology associated with DNA sequences.

Journals no longer print full sequence data, but instead print a database accession number, and require authors to submit sequences to a public database when an article describing a new sequence is submitted for publication. Many scientists release their sequences before the article detailing them is in press. This practice is now the rule for large genome centers, and, although some individual laboratories still wait for acceptance of publication before making their data available, others consider the release of a record to be publication in its own right.

The submission process is governed by an international, collaborative agreement. Sequences submitted to any one of the three databases participating in this collaboration will appear in the other two databases within a few days of their release to the public. Sequence records are then distributed worldwide by various user groups and centers, including those that reformat the records for use within their own suites of programs and databases. Thus, by submitting a sequence to only one of the three ''major'' databases, researchers can quickly disseminate their sequence data and avoid the possibility that redundant records will be archived.

As mentioned often in this book, the growth of sequence databases has been exponential. Most sequence records in the early years were submitted by individual scientists studying a gene of interest. A program suitable for this type of submission should allow for the manual annotation of arbitrary biological information. However, the databases recently have had to adapt not only to new classes of data but also to a substantially higher rate of submission. A significant fraction of submissions now represents phylogenetic and population studies, in which relationships between sequences need to be explicitly demonstrated. Completed genomes are also becoming available at a growing rate.

This chapter is devoted to the submission of DNA and protein sequences and their annotations into the public databases. Presented here are two different approaches for submitting sequences to the databases, one Web-based (using BankIt) and the other using Sequin, a multi-platform program that can use a direct network connection. Sequin is also an ASN.1 editing tool that takes full advantage of the NCBI data model (see Chapter 2) and has become a platform for many sequence analysis tools that NCBI has developed over the years. (A separate bulk-submission protocol used for EST records, which are submitted to the databases at the rate of thousands per day, is discussed briefly at the end of this chapter. Fortunately, EST records are fairly simple and uniform in content, making them amenable to automatic processing.)

WHY, WHERE, AND WHAT TO SUBMIT?

One should submit to whichever of the three public databases is most convenient. This may be the database that is closest geographically, it may be the repository one has always used in the past, or it may simply be the place one's submission is likely to receive the best attention. All three databases have knowledgeable staff able to help submitters throughout the process. Under normal circumstances, an accession number will be returned within one workday, and a finished record should be available within 5–10 working days, depending on the information provided by the submitter. Submitting data to the database is not the end of one's scientific obligation. Updating the record as more information becomes available will ensure that the information within the record will survive time and scientific rigor.

Presently, it is assumed that all submissions of sequences are done electronically: via the World Wide Web, by electronic mail, or (at the very least) on a computer disk sent via regular postal mail. The URLs and E-mail addresses for electronic submissions are shown in the list at the end of the chapter.

All three databases want the same end result: a richly annotated, biologically and computationally sound record, one that allows other scientists to be able to reap the benefits of the work already performed by the submitting biologist and that affords links to the protein, bibliographic, and genomic databases (see Chapter 7). There is a rich set of biological features and other annotations available, but the important components are the ones that lend themselves to analysis. These include the nucleotide and protein sequences, the CDS (coding sequence, also known as coding region), gene, and mRNA features (i.e., features representing the central dogma of molecular biology), the organism from which the sequences were determined, and the bibliographic citation that links them to the information sphere and will have all the experimental details that give this sequence its *raison d'être.*

DNA/RNA

The submission process is quite simple, but care must be taken to provide information that is accurate (free of errors and vector or mitochondrial contamination) and as biologically sound as possible, to ensure maximal usability by the scientific community. Here are a few matters to consider before starting a submission, regardless of its form.

Nature of the Sequence. Is it of genomic or mRNA origin? Users of the databases like to know the nature of the physical DNA that is the origin of the molecule being sequenced. For example, although cDNA sequencing is performed on DNA (and not RNA), the type of the molecule present in the cell is mRNA. The same is true for the genomic sequencing of rRNA genes, in which the sequenced molecule is almost always genomic DNA. Copying the rRNA into DNA, like direct sequencing of rRNA, although possible, is rarely done. Bear in mind also that, because the sequence being submitted should be of a unique molecular type, it must not represent (for example) a mixture of genomic and mRNA molecule types that cannot actually be isolated from a living cell.

Is the Sequence Synthetic, But Not Artificial? There is a special division in the nucleotide databases for synthetic molecules, sequences put together experimentally that do not occur naturally in the environment (e.g., protein expression vector sequences). The DNA sequence databases do not accept computer-generated sequences, such as consensus sequences, and all sequences in the databases are experimentally derived from the actual sequencing of the molecule in question. They can, however, be the compilation of a shotgun sequencing exercise.

How Accurate is the Sequence? This question is poorly documented in the database literature, but the assumption that the submitted sequence is as accurate as possible usually means at least two-pass coverage (in opposite orientations) on the whole submitted sequence. Equally important is the verification of the final submitted sequence. It should be free of vector contamination (this can be verified with a BLASTN search against the VecScreen database; see Chapter 8 and later in this

chapter) and possibly checked with known restriction maps, to eliminate the possibility of sequence rearrangement and to confirm correct sequence assembly.

Organism

All DNA sequence records must show the organism from which the sequence was derived. Many inferences are made from the phylogenetic position of the records present in the databases. If these are wrongly placed, an incorrect genetic code may be used for translation, with the possible consequence of an incorrectly translated or prematurely truncated protein product sequence. Just knowing the genus and species is usually enough to permit the database staff to identify the organism and its lineage. NCBI offers an important taxonomy service, and the staff taxonomists maintain the taxonomy that is used by all the nucleotide databases and by SWISS-PROT, a curated protein database.

Citation

As good as the annotations can be, they will never surpass a published article in fully representing the state of biological knowledge with respect to the sequence in any given record. It is therefore imperative to ensure the proper link between the research publication and the primary data it will cite. For this reason, having a citation in the submission being prepared is of great importance, even if it consists of just a temporary list of authors and a working title. Updating these citations at publication time is also important to the value of the record. (This is done routinely by the database staff and will happen more promptly if the submitter notifies the staff on publication of the article.)

Coding Sequence(s)

A submission of nucleotide also means the inclusion of the protein sequences it encodes. This is important for two reasons:

- Protein databases (e.g., SWISS-PROT and PIR) are almost entirely populated by protein sequences present in DNA sequence database records.
- The inclusion of the protein sequence serves as an important, if not essential, validation step in the submission process.

Proteins include the enzyme molecules that carry out many of the biological reactions we study, and their sequences are an intrinsic part of the submission process. Their importance, which is discussed in Chapter 2, is also reflected in the submission process, and this information must be captured for representation in the various databases. Also important are the protein product and gene names, if these are known. There are a variety of resources (many present in the lists that conclude these chapters) that offer the correct gene nomenclature for many organisms (cf. Genetic nomenclature guide, *Trends in Genetics*, 1998).

The coding sequence features, or CDS, are the links between the DNA or RNA and the protein sequences, and their correct positioning is central in the validation, as is the correct genetic code. The nucleotide databases now use 17 different genetic

codes that are maintained by the taxonomy and molecular biology staff at NCBI. Because protein sequences are so important, comprising one of the main pieces of molecular biology information on which biologists can compute, they receive much deserved attention from the staff at the various databases. It is usually simple to find the correct open-reading frame in an mRNA (see Chapter 10), and various tools are available for this (e.g., NCBI's ORF Finder). Getting the correct CDS intervals in a genomic sequence from a higher eukaryote is a little trickier: the different exon-coding sequences must be joined, and this involves a variety of approaches, also described in Chapter 10. (The Suggest Intervals function in Sequin will calculate CDS intervals if given the sequence of the protein and the proper genetic code.) A submitted record will be validated by the database staff but even more immediately by the submission tool used as well. Validation checks that the start and stop codons are included in the CDS intervals, that these intervals are using exon/intron-consensus boundaries, and that the provided amino acid sequence can be translated from the designated CDS intervals using the appropriate genetic code.

Other Features

There are a variety of other features available for the feature sections of a submitted sequence record. The complete set of these is represented in the feature table documentation. Although many features are available, there is much inconsistent usage in the databases, mainly due to a lack of consistent guidelines and poor agreement among biologists as to what they really mean. Getting the organism, bibliography, gene, CDS, and mRNA correct usually suffices and makes for a record that can be validated, is informative, and allows a biologist to grasp in a few lines of text an overview of the biology of the sequence. Nonetheless, the full renditions of the feature table documentation are available for use as appropriate but with care taken as to the intent of the annotations.

POPULATION, PHYLOGENETIC, AND MUTATION STUDIES

The nucleotide databases are now accepting population, phylogenetic, and mutational studies as submitted sequence sets, and, although this information is not adequately represented in the flatfile records, it is appearing in the various databases. This allows the submission of a group of related sequences together, with entry of shared information required only once. Sequin also allows the user to include the alignment generated with a favorite alignment tool and to submit this information with the DNA sequence. New ways to display this information (such as Entrez) should soon make this kind of data more visible to the general scientific community.

PROTEIN-ONLY SUBMISSIONS

In most cases, protein sequences come with a DNA sequence. There are some exceptions—people do sequence proteins directly—and such sequences must be submitted without a corresponding DNA sequence. SWISS-PROT presently is the best venue for these submissions.

HOW TO SUBMIT ON THE WORLD WIDE WEB

The World Wide Web is now the most common interface used to submit sequences to the three databases. The Web-based submission systems include Sakura (''cherry blossoms'') at DDBJ, WebIn at EBI, and BankIt at the NCBI. The Web is the preferred submission path for simple submissions or for those that do not require complicated annotations or too much repetition (i.e., 30 similar sequences, as typically found in a population study, would best be done with Sequin, see below). The Web form is ideal for a research group that makes few sequence submissions and needs something simple, entailing a short learning curve. The Web forms are more than adequate for the majority of the submissions: some 75–80% of individual submissions to NCBI are done via the Web. The alternative addresses (or URLs) for submitting to the three databases are presented in the list at the end of the chapter.

On entering a BankIt submission, the user is asked about the length of the nucleotide sequence to be submitted. The next BankIt form is straightforward: it asks about the contact person (the individual to whom the database staff may address any questions), the citations (who gets the scientific credit), the organism (the top 100 organisms are on the form; all others must be typed in), the location (nuclear vs. organelle), some map information, and the nucleotide sequence itself. At the end of the form, there is a BankIt button, which calls up the next form. At this point, some validation is made, and, if any necessary fields were not filled in, the form is presented again. If all is well, the next form asks how many features are to be added and prompts the user to indicate their types. If no features were added, BankIt will issue a warning and ask for confirmation that not even one CDS is to be added to the submission. The user can say no (zero new CDSs) or take the opportunity to add one or more CDS. At this point, structural RNA information or any other legal DDBJ/EMBL/GenBank features can be added as well.

To begin to save a record, press the BankIt button again. The view that now appears must be approved before the submission is completed; that is, more changes may be made, or other features may be added. To finish, press BankIt one more time. The final screen will then appear; after the user toggles the Update/Finished set of buttons and hits BankIt one last time, the submission will go to NCBI for processing. A copy of the just-finished submission should arrive promptly via E-mail; if not, one should contact the database to confirm receipt of the submission and to make any correction that may be necessary.

HOW TO SUBMIT WITH SEQUIN

Sequin is designed for preparing new sequence records and updating existing records for submission to DDBJ, EMBL, and GenBank. It is a tool that works on most computer platforms and is suitable for a wide range of sequence lengths and complexities, including traditional (gene-sized) nucleotide sequences, segmented entries (e.g., genomic sequences of a spliced gene for which not all intronic sequences have been determined), long (genome-sized) sequences with many annotated features, and sets of related sequences (i.e., population, phylogenetic, or mutation studies of a particular gene, region, or viral genome). Many of these submissions could be performed via the Web, but Sequin is more practical for more complex cases. Certain

types of submission (e.g., segmented sets) cannot be made via the Web unless explicit instructions to the database staff are inserted.

Sequin also accepts sequences of proteins encoded by the submitted nucleotide sequences and allows annotation of features on these proteins (e.g., signal peptides, transmembrane regions, or cysteine disulfide bonds). For sets of related or similar sequences (e.g., population or phylogenetic studies), Sequin accepts information from the submitter on how the multiple sequences are aligned to each other. Finally, Sequin can be used to edit and resubmit a record that already exists in GenBank, either by extending (or replacing) the sequence or by annotating additional features or alignments.

Submission Made Easy

Sequin has a number of attributes that greatly simplify the process of building and annotating a record. The most profound aspect is automatic calculation of the intervals on a CDS feature given only the nucleotide sequence, the sequence of the protein product, and the genetic code (which is itself automatically obtained from the organism name). This ''Suggest Intervals'' process takes consensus splice sites into account in its calculations. Traditionally, these intervals were entered manually, a time-consuming and error-prone process, especially on a genomic sequence with many exons, in cases of alternative splicing, or on segmented sequences.

Another important attribute is the ability to enter relevant annotation in a simple format in the definition line of the sequence data file. Sequin recognizes and extracts this information when reading the sequences and then puts it in the proper places in the record. For nucleotide sequences, it is possible to enter the organism's scientific name, the strain or clone name, and several other source modifiers. For example

```
>eIF4E [organism=Drosophila melanogaster] [strain=Oregon R]
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGTTCAACAATCGATA
GCTGCCTTTGGCCACCAAAATCCCAAACTTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCCAG
...
```

This is especially important for population and phylogenetic studies, where the source modifiers are necessary to distinguish one component from another.

For protein sequences, the gene and protein names can be entered. For example

```
>4E-I [gene=eIF4E] [protein=eukaryotic initiation factor 4E-I]
MQSDFHRMKNFANPKSMFKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGNTATTTAPAGD
DAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTVEDFWSLYNHIKPPSEIKLGSDYSLFKK
...
```

If this information is not present in the sequence definition line, Sequin will prompt the user for it before proceeding. Annotations on the definition line can be very convenient, since the information stays with the sequence and cannot be forgotten or mixed-up later. In addition to building the proper CDS feature, Sequin will automatically make gene and protein features with this information.

Because the majority of submissions contain a single nucleotide sequence and one or more coding region features (and their associated protein sequences), the functionality just outlined can frequently result in a finished record, ready to submit

without any further annotation. With gene and protein names properly recorded, the record becomes informative to other scientists who may retrieve it as a BLAST similarity result or from an Entrez search.

Starting a New Submission

Sequin begins with a window that allows the user to start a new submission or load a file containing a saved record. After the initial submission has been built, the record can be saved to a file and edited later, before finally being sent to the database. If Sequin has been configured to be network aware, this window also allows the downloading of existing database records that are to be updated.

A new submission is made by filling out several forms. The forms use folder tabs to subdivide a window into several pages, allowing all the requested data to be entered without the need for a huge computer screen. These entry forms have buttons for Prev(ious) Page and Next Page. When the user arrives at the last page on a form, the Next Page button changes to Next Form.

The Submitting Authors form requests a tentative title, information on the contact person, the authors of the sequence, and their institutional affiliations. This form is common to all submissions, and the contact, authors, and affiliation page data can be saved by means of the Export menu item. The resulting file can be read in when starting other submissions by choosing the Import menu item. However, because even population, phylogenetic, or mutation studies are submitted in one step as one record, there is less need to save the submitter information.

The Sequence Format form asks for the type of submission (single sequence, segmented sequence, or population, phylogenetic, or mutation study). For the last three types of submission, which involve comparative studies on related sequences, the format in which the data will be entered also can be indicated. The default is FASTA format (or raw sequence), but various contiguous and interleaved formats (e.g., PHYLIP, NEXUS, PAUP, and FASTA+ GAP) are also supported. These latter formats contain alignment information, and this is stored in the sequence record.

The Organism and Sequences form asks for the biological data. On the Organism page, as the user starts to type the scientific name, the list of frequently used organisms scrolls automatically. (Sequin holds information on the top 800 organisms present in GenBank.) Thus, after typing a few letters, the user can fill in the rest of the organism name by clicking on the appropriate item in the list. Sequin now knows the scientific name, common name, GenBank division, taxonomic lineage, and, most importantly, the genetic code to use. (For mitochondrial genes, there is a control to indicate that the alternative genetic code should be used.) For organisms not on the list, it may be necessary to set the genetic code control manually. Sequin uses the standard code as the default. The remainder of the Organism and Sequences form differs depending on the type of submission.

Entering a Single Nucleotide Sequence and its Protein Products

For a single sequence or a segmented sequence, the rest of the Organism and Sequences form contains Nucleotide and Protein folder tabs. The Nucleotide page has controls for setting the molecule type (e.g., genomic DNA or mRNA) and topology (usually linear, occasionally circular) and for indicating whether the sequence is

incomplete at the 5′ or 3′ ends. Similarly, the Protein page has controls for creating an initial mRNA feature and for indicating whether the sequence is incomplete at the amino or carboxyl ends.

For each protein sequence, Suggest Intervals is run against the nucleotide sequence (using the entered genetic code, which is usually deduced from the chosen organism), and a CDS feature is made with the resulting intervals. A Gene feature is generated, with a single interval spanning the CDS intervals. A protein product sequence is made, with a Protein feature to give it a name. The organism and publication are placed so as to apply to all nucleotide and protein sequences within the record. Appropriate molecule-type information is also placed on the sequences. In most cases, it is much easier to enter the protein sequence and let Sequin construct the record automatically than to manually add a CDS feature (and associated gene and protein features) later.

Entering an Aligned Set of Sequences

A growing class of submissions involves sets of related sequences: population, phylogenetic, or mutation studies. A large number of HIV sequences come in as population studies. A common phylogenetic study involves ribulose-1,5-bisphosphate carboxylase (RUBISCO), a major enzyme of photosynthesis and perhaps the most prevalent protein (by weight) on earth. Submitting such a set of sequences is not much more complex than submitting a single sequence. The same submission information form is used to enter author and contact information.

In the Sequence Format form, the user chooses the desired type of submission. Population studies are generally from different individuals in the same (cross-breeding) species. Phylogenetic studies are from different species. In the former case, it is best to embed in the definition lines strain, clone, isolate, or other source-identifying information. In the latter case, the organism's scientific name should be embedded. Multiple sequence studies can be submitted in FASTA format, in which case Sequin should later be called on to calculate an alignment. Better yet, alignment information can be indicated by encoding the data in one of several popular alignment formats.

The Organism and Sequences form is slightly different for sets of sequences. The Organism page for phylogenetic studies allows the setting of a default genetic code only for organisms not in Sequin's local list of popular species. The Nucleotide page has the same controls as for a single sequence submission. Instead of a Protein page, there is now an Annotation page. Many submissions are of rRNA sequence or no more than a complete CDS. (This means that the feature intervals span the full range of each sequence.) The Annotation page allows these to be created and named. A definition line (title) can be specified, and Sequin can prefix the individual organism name to the title. More complex situations, in which sequences have more than a single interval feature across the entire span, can be annotated by feature propagation after the initial record has been built and one of the sequences has been annotated.

As a final step, Sequin displays an editor that allows all organism and source modifiers on each sequence to be edited (or entered if the definition lines were not annotated). On confirmation of the modifiers, Sequin finishes assembling the record into the proper structure.

Viewing the Sequence Record

Sequin provides a number of different views of a sequence record. The traditional flatfile can be presented in FASTA, GenBank (Fig. 4.1), or EMBL format. (These can be exported to files on the user's computer, which can then be entered into other sequence analysis packages.) A graphical view (Fig. 4.2) shows feature intervals on a sequence. This is particularly useful for viewing alternatively spliced coding regions. (The style of the Graphical view can be customized, and these views can also be copied to the personal computer's clipboard for pasting into a word processor or drawing program that will be used in preparing a manuscript for publication.) There is a more detailed view that shows the features on the actual sequence. For records containing alignments (e.g., alignments between related sequences entered by a user, or the results of a BLAST search), one can request either a graphical

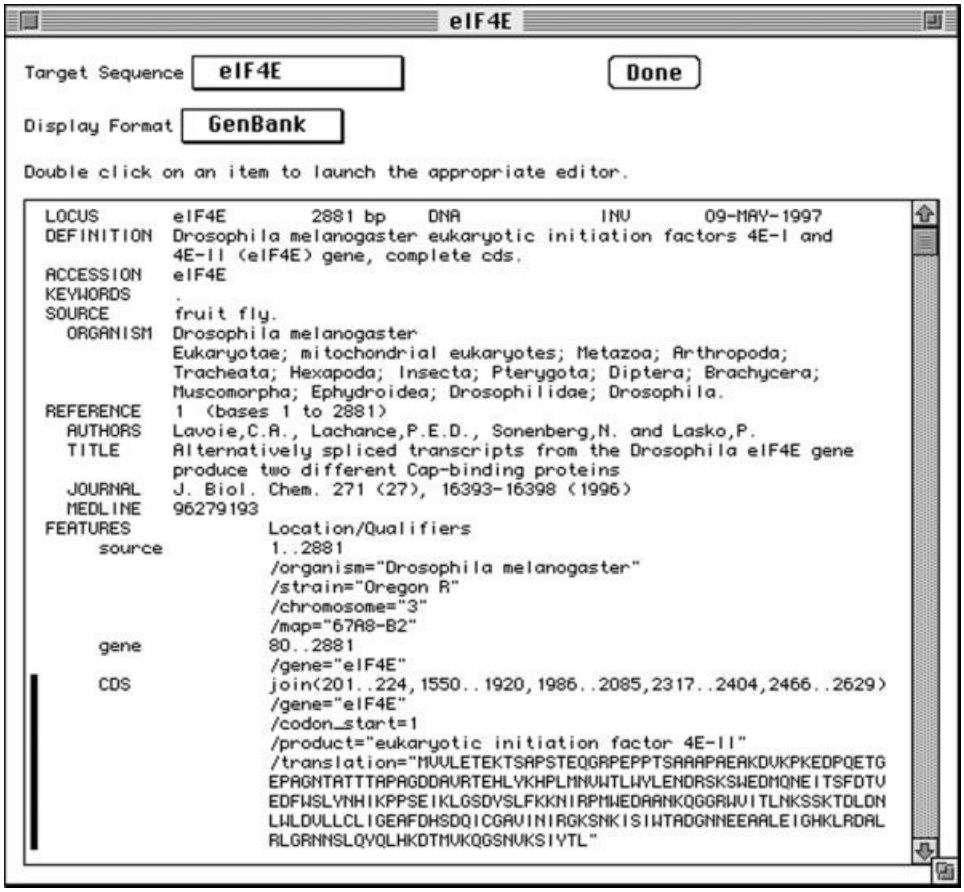

Figure 4.1. Viewing a sequence record with Sequin. The sequence record viewer uses GenBank format, by default. In this example, a CDS feature has been clicked, as indicated by the bar next to its paragraph. Double-clicking on a paragraph will launch an editor for the feature, descriptor, or sequence that was selected. The viewer can be duplicated, and multiple viewers can show the same record in different formats.

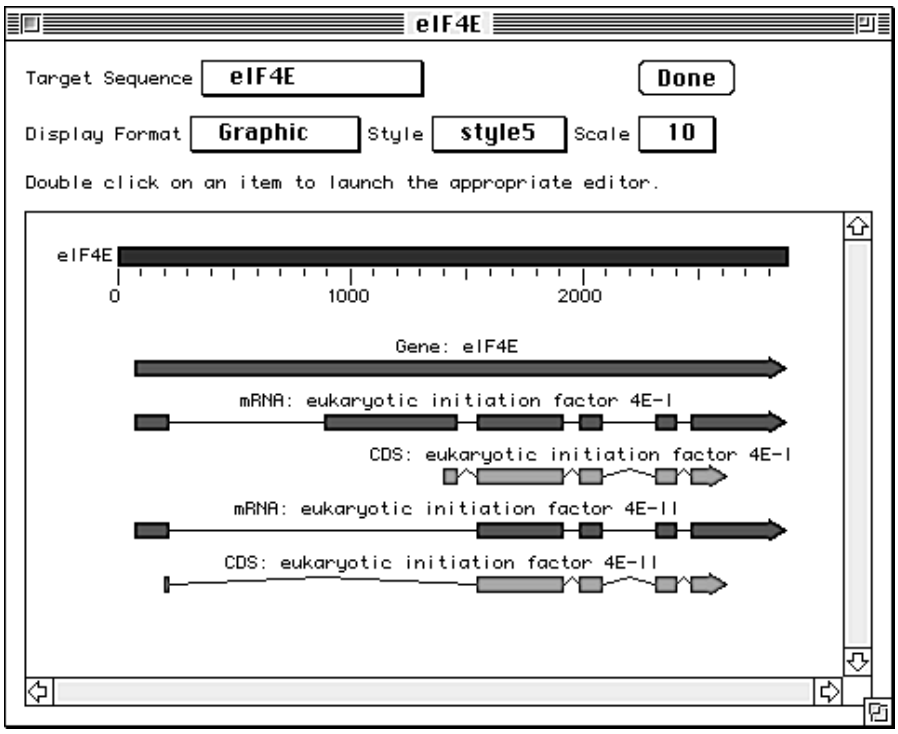

Figure 4.2. Sequin's graphical format can show segmented sequence construction and feature intervals. These can be compared with drawings in laboratory notebooks to see, at a glance, whether the features are annotated at the proper locations. Different styles can be used, and new styles can be created, to customize the appearance of the graphical view. The picture can be copied to a personal computer's clipboard for pasting into a word processor or drawing program.

overview showing insertions, deletions, and mismatches or a detailed view showing the alignment of sequence letters.

The above-mentioned viewers are interactive. Clicking on a feature, a sequence, or the graphical representation of an alignment between sequences will highlight that object. Double-clicking will launch the appropriate editor. Multiple viewers can be used on the same record, permitting different formats to be seen simultaneously. For example, it is quite convenient to have the graphical view and the GenBank (or EMBL) flatfile view present at the same time, especially on larger records containing more than one CDS. The graphical view can be compared to a scientist's lab notebook drawings, providing a quick reality check on the overall accuracy of the feature annotation.

Validation

To ensure the quality of data being submitted, Sequin has a built-in validator that searches for missing organism information, incorrect coding region lengths (compared to the submitted protein sequence), internal stop codons in coding regions,

mismatched amino acids, and nonconsensus splice sites. Double-clicking on an item in the error report launches an editor on the ''offending'' feature.

The validator also checks for inconsistent use of ''partial'' indications, especially among coding regions, the protein product, and the protein feature on the product. For example, if the coding region is marked as incomplete at the 5′ end, the protein product and protein feature should be marked as incomplete at the amino end. (Unless told otherwise, the CDS editor will automatically synchronize these separate partial indicators, facilitating the correction of this kind of inconsistency.)

Advanced Annotation and Editing Functions

The sequence editor built into Sequin automatically adjusts feature intervals as the sequence is edited. This is particularly important if one is extending an existing record by adding new 5′ sequence. Prior to Sequin, this process entailed manually correcting the intervals on all biological features on the sequence or, more likely, redoing the entire submission from scratch. The sequence editor is used much like a text editor, with new sequence being pasted in or typed in at the position of a cursor.

For population or phylogenetic studies, Sequin allows annotation of one sequence, whereupon features from that sequence can be propagated to all other sequences through the supplied alignment. (In the case of a CDS feature, the feature intervals can be calculated automatically by reading in the sequence of its protein product rather than having to enter them by typing.) Feature propagation is accessed from the alignment editor. The result is the same as would have been achieved if features had been manually annotated on each sequence, but with feature propagation the entire process can be completed in minutes rather than hours.

The concepts behind feature propagation and the sequence editor combine to provide a simple and automatic method for updating an existing sequence. The Update Sequence functions allow the user to enter an overlapping sequence or a replacement sequence. Sequin makes an alignment, merges the sequences if necessary, propagates features onto the new sequence in their new positions, and uses these to replace the old sequence and features.

Genome centers frequently store feature coordinates in databases. Sequin can now annotate features by reading a simple tab-delimited file that specifies the location and type of each feature. The first line starts with `>Features`, a space, and the sequence identifier of the sequence. The table is composed of five columns: start, stop, feature key, qualifier key, and qualifier value. The columns are separated by tab characters. The first row for any given feature has start, stop, and feature key. Additional feature intervals just have start and stop. The qualifiers follow on lines starting with three tabs. An example of this format follows below.

```
>Features lcl|eIF4E
80      2881      gene
                        gene      eIF4E
1402    1458      CDS
1550    1920
1986    2085
2317    2404
2466    2629
                        product   eukaryotic initiation factor 4E-I
```

Sending the Submission

A finished submission can be saved to disk and E-mailed to one of the databases. It is also a good practice to save frequently throughout the Sequin session, to make sure nothing is inadvertently lost. The list at the end of this chapter provides E-mail addresses and contact information for the three databases.

UPDATES

The database staffs at all three databases welcome all suggestions on making the update process as efficient and painless as possible. People who notice that records are published but not yet released are strongly encouraged to notify the databases as well. If errors are detected, these should also be forwarded to the updates addresses; the owner of the record is notified accordingly (by the database staff), and a correction usually results. This chain of events is to be distinguished from third-party annotations, which are presently not accepted by the databases. *The record belongs to the submitter(s)*; the database staff offers some curatorial, formatting guideline suggestions, but substantive changes come only from a listed submitter. Many scientists simply E-mail a newly extended sequence or feature update to the databases for updating.

CONSEQUENCES OF THE DATA MODEL

Sequin is, in reality, an ASN.1 editor. The NCBI data model, written in the ASN.1 data description language, is designed to keep associated information together in descriptors or features (see Chapter 2). Features are typically biological entities (e.g., genes, coding regions, RNAs, proteins) that always have a location (of one or more intervals) on a sequence. Descriptors were introduced to carry information that can apply to multiple sequences, eliminating the need to enter multiple copies of the same information.

For the simplest case, that of a single nucleotide sequence with one or more protein products, Sequin generally allows the user to work without needing to be aware of the data model's structural hierarchy. Navigation is necessary, as is at least a cursory understanding of the data model, if extensive annotation on protein product sequences is contemplated or for manual annotation of population and phylogenetic sets. Setting the Target control to a given sequence changes the viewer to show a graphical view or text report on that sequence. Any features or descriptors created with the Annotation submenus will be packaged on the currently targeted sequence.

Although Sequin does provide full navigation among all sequences within a structured record, building the original structure from the raw sequence data is a job best left to Sequin's ''create new submission'' functions described above. Sequin asks up front for information (e.g., organism and source modifiers, gene and protein names) and knows how to correctly package everything into the appropriate place. This was, in fact, one of the main design goals of Sequin. Manual annotation requires a more detailed understanding of the data model and expertise with the more esoteric functions of Sequin.

Using Sequin as a Workbench

Sequin also provides a number of sequence analysis functions. For example, one function will reverse-complement the sequence and the intervals of its features. New functions can easily be added. These functions appear in a window called the NCBI Desktop (Fig. 4.3), which directly displays the internal structure of the records currently loaded in memory. This window can be understood as a Venn diagram, with descriptors on a set (such as a population study) applying to all sequences in that set. The Desktop allows drag-and-drop of items within a record. For example, the

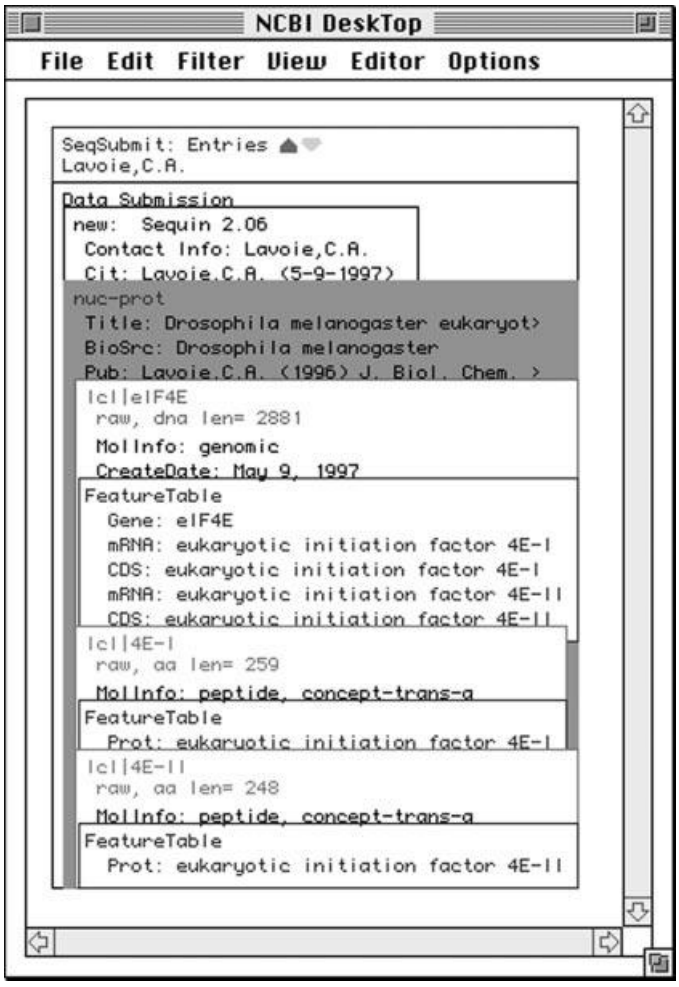

Figure 4.3. The NCBI Desktop displays a graphical overview of how the record is structured in memory, based on the NCBI data model (see Chapter 2). This view is most useful to a software developer or database sequence annotator. In this example, the submission contains a single Nuc-prot set, which in turn contains a nucleotide and two proteins. Each sequence has features associated with it. BioSource and publication descriptors on the Nuc-prot set apply the same organism (*Drosophila melanogaster*) and the same publication, respectively, to all sequences.

user can read the results of a BLAST analysis and then drag-and-drop this information onto a sequence record, thus adding the alignment data to the record, or a newly calculated feature can be dragged into the record. (A separate Seq-loc on the Desktop can be dragged onto a feature, in which case it changes the feature location.) The modifications are immediately displayed on any active viewers. Note, however, that not all annotations are visible in all viewers. The flatfile view does have its limitations; for example, it does not display alignments and does not even indicate that alignments are present. Understanding the Desktop is not necessary for the casual user submitting a simple sequence; however, for the advanced user, it can immediately take the mystery out of the data.

EST/STS/GSS/HTG/SNP AND GENOME CENTERS

Genome centers have now established a number of relationships with DNA sequence databases and have streamlined the submission process for a great number of record types. Not all genome centers deal with all sequence types, but all databases do. The databases have educated their more sophisticated users on this, and, conversely, some of the genomes centers have also encouraged certain database managers to learn their own data model as well (e.g., the use of AceDB to submit sequences at Stanford, Washington University at St. Louis, and the Sanger Centre or the use of XML at Celera).

CONCLUDING REMARKS

The act of depositing records into a database and seeing these records made public has always been an exercise of pride on the part of submitters, a segment of the scientific activity from their laboratory that they present to the scientific community. It is also a mandatory step that has been imposed by publishers as part of the publication process. In this process, submitters always hope to provide information in the most complete and useful fashion, allowing maximum use of their data by the scientific community.

Very few users are aware of the complete array of intricacies present in the databases, but they do know the biology they want these entries to represent. It is incumbent on the databases to provide tools that will facilitate this process. The database staff also provides expertise through their indexing staff (some databases also call them *curators* or *annotators*), who have extensive training in biology and are very familiar with the databases; they ensure that nothing is lost in the submission process. The submission exercise itself has not always been easy and was not even encouraged at the beginning of the sequencing era, simply because databases did not know how to handle this information. Now, however, the databases strongly encourage the submission of sequence data and of all appropriate updates. Many tools are available to facilitate this task, and together the databases support Sequin as the tool to use for new submissions, in addition to their respective Web submissions tools. Submitting data to the databases has now become a manageable (and sometimes enjoyable) task, with scientists no longer having good excuses for neglecting it.

CONTACT POINTS FOR SUBMISSION OF SEQUENCE DATA TO DDBJ/EMBL/GenBank

DDBJ (Center for Information Biology, NIG)

Address	DDBJ, 1111 Yata, Mishima, Shiznoka 411, Japan
Fax	81-559-81-6849
E-mail	
Submissions	*ddbjsub@ddbj.nig.ac.jp*
Updates	*ddbjupdt@ddbj.nig.ac.jp*
Information	*ddbj@ddbj.nig.ac.jp*
World Wide Web	
Home page	*http://www.ddbj.nig.ac.jp/*
Submissions	*http://sakura.ddbj.nig.ac.jp/*

EMBL (European Bioinformatics Institutes, EMBL Outstation)

Address	EMBL Outstation, EBI, Wellcome Trust Genome Campus, Hinxton Cambridge, CB10 1SD, United Kingdom
Voice	01.22.349.44.44
Fax	01.22.349.44.68
E-mail	
Submissions	*datasubs@ebi.ac.uk*
Updates	*update@ebi.ac.uk*
Information	*datalib@ebi.ac.uk*
World Wide Web	
Home page	*http://www.ebi.ac.uk/*
Submissions	*http://www.ebi.ac.uk/subs/allsubs.html*
WebIn	*http://www.ebi.ac.uk/submission/webin.html*

GenBank (National Center for Biotechnology Information, NIH)

Address	GenBank, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N805, Bethesda MD 20894
Telephone	301-496-2475
Fax	301-480-9241
E-mail	
Submissions	*gb-sub@ncbi.nlm.nih.gov*
EST/GSS/STS	*batch-sub@ncbi.nlm.nih.gov*
Updates	*update@ncbi.nlm.nih.gov*
Information	*info@ncbi.nlm.nih.gov*
World Wide Web	
Home page	*http://www.ncbi.nlm.nih.gov/*
Submissions	*http://www.ncbi.nlm.nih.gov/Web/GenBank/submit.html*
BankIt	*http://www.ncbi.nlm.nih.gov/BankIt/*

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 4

dbEST	*http://www.ncbi.nlm.nih.gov/dbEST/*
dbSTS	*http://www.ncbi.nlm.nih.gov/dbSTS/*
dbGSS	*http://www.ncbi.nlm.nih.gov/dbGSS/*

DDBJ/EMBL/GenBank Feature Table Documentation	*http://www.ncbi.nlm.nih.gov/collab/FT/*
EMBL Release Notes	*ftp://ftp.ebi.ac.uk/pub/databases/embl/release/relnotes.doc*
GenBank Release Notes	*ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt*
Genetic codes used in DNA sequence databases	*http://www.ncbi.nlm.nih.gov/htbinpost/Taxonomy/wprintgc?mode_c*
HTGS	*http://www.ncbi.nlm.nih.gov/HTGS/*
ORF Finder	*http://www.ncbi.nlm.nih.gov/gorf/gorf.html*
Sequin	*http://www.ncbi.nlm.nih.gov/Sequin/*
Taxonomy browser	*http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html*

REFERENCES

Boguski, M. S., Lowe, T. M., Tolstoshev, C. M. (1993). dbEST—database for ''expressed sequence tags. *Nat. Genet.* 4, 332–333.

Ouellette, B. F. F., and Boguski, M. S. 1997. Database divisions and homology search files: a guide for the perplexed. *Genome Res.* 7, 952–955.